DOCUMENT RESUME

ED 137 384 95 TM 006 190

AUTHOR Powell, Marjorie
TITLE Necessary Steps to Insure Availability of Data for Secondary Analysis.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE [Apr 77]
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Civil Liberties; *Data; *Data Analysis; *Data Bases; *Data Collection; Information Dissemination; Information Needs; Information Sources; *Research Needs
IDENTIFIERS *Secondary Data Analysis

ABSTRACT

The increasing cost of data collection, increasing complexity of data analysis procedures, and limited funding for educational research require greater utilization of existing data sets for multiple purposes. At the same time, the problems associated with such access to data require the development of guidelines by the profession which will both alert legislators and regulatory agencies to the problems and concerns of the profession and at the same time provide assistance to members of the profession as they confront issues related to, and specific assistance of, requests for access to data for secondary analysis. Recommendations include those for persons beginning data collection efforts to insure that data can be used for secondary analysis, and for the profession to provide guidelines and mechanisms to make data available. The need for discussion, to clarify all sides of several issues, is emphasized. The establishment of a data bank might serve as a catalyst for such discussion. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED137384

NECESSARY STEPS TO INSURE AVAILIBILITY OF DATA FOR SECONDARY ANALYSIS

Marjorie Powell

California Commission for Teacher Preparation and Licensing

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM006 190

A paper presented at the meetings of the American Educational Research Association, New York City, New York, April 4-8, 1977.

The development of this paper was supported, in part, by the Beginning Teacher Evaluation Study, conducted by the California Commission for Teacher Preparation and Licensing with funds provided by the National Institute of Education. However, the opinions expressed are those of the author and do not represent an official position of the California Commission for Teacher Preparation and Licensing.

A number of reasons have been advanced in support of the use of existing data sets for further analysis. At the same time, there are a number of problems facing any individual or agency which wishes either to make data available or to use data collected by another person or agency. While a number of individuals and agencies may be forced to address these issues independently, in many cases the issues are so large as to preclude adequate solution by an individual or single agency. In other cases, agencies are covering ground which has been covered by others, frequently without the advice and assistance of those others. We each start out with a unique and separate problem and attempt a solution within the limits of our skills and resources, without benefit of the knowledge and procedures acquired by others.

It is time that the profession as a whole addressed some of these problems, and at the very least provide for extensive discussion to illuminate the many sides of each concern.

What are these problems?

Collection of Data

The rules and regulations relating to privacy and the protection of individuals, as well as access of researchers to data normally maintained by various agencies, have an impact upon the extent to which data can be collected at all, and the extent to which data collected for one purpose can be made available to other researchers for other purposes. Procedures must be used which will protect the rights to privacy of the individuals who participate in research efforts while allowing the data to be

used for a number of purposes. The public, which finances the vast majority of data collection through the taxes which support federally funded research efforts, has the right to thorough and efficient use of the data collected by those research efforts. The public should not be asked to pay for the collection of data when an existing data set can be used to answer the research questions of interest.

Release of Data

In most instances the principal investigator is responsible for decisions about what data to release, to whom, under what circumstances. However, a number of forces appear to be pushing toward the removal of such rights and responsibilities from the principal investigator and their assignment to other persons or groups. When studies are conducted specifically to inform policy or to evaluate programs, there may be a desire for multiple analyses of data sets to more fully elucidate the questions of interest. As the cost of data collection increases, more thorough utilization of collected data will become necessary.

However, principal investigator or committee, many of the persons making such decisions have limited familiarity with possible procedures for modifying data sets to prevent the reconstruction of the sample. Discussion of such procedures is beyond the scope of this paper, but the interested reader is referred to several articles by Dr. Robert Boruch.

Aside from questions of modifying or disguising data sets, there is wide variation in the criteria used in reviewing data for

release. Within one research project, the Beginning Teacher Evaluation Study, one principal investigator has indicated a willingness to release all data without names, while another principal investigator has removed any data which, in conjunction with other data, might enable a third party to reconstruct the sample. Estimates of the amount of data thus removed from archival data tapes range from 40 to 60 percent of the original data set. Obviously, very different criteria are being applied in these two instances.

At present, each person or agency considering the release of data for further analyses must develop a set of criteria to be used to determine what data to release. This is often done with limited knowledge of criteria used by other agencies. Confusion, misunderstandings and differing limits on the usability of various data sets may result from the use of different criteria.

The persons and agencies involved in the collection and release of data need protection from misuse of the data by persons conducting further analysis. Aside from the deliberate misuse of data, problems may result from analyses, by the secondary analyst, which the primary researcher promised participants would not be conducted. For example, participants in the BTES were promised that no direct comparisons of schools or districts would be made. I would consider that any further analysis of these data which made such direct comparisons would be a misuse of the data.

Protection of the persons and agencies collecting data also involves protection of the right of the investigator(s) to conduct

analyses of their data and publish the results of their analyses. When the research is conducted with report delivery dates, there may be a number of other analyses that the primary researchers wish to conduct at the completion of the contracted report.

At one end of the continuum, all data should be released immediately for any further analyses. However, misuse of data may result in harm to the data collectors, from loss of volunteer participants to lawsuits. On the other hand, if data are released only for some purposes, decisions have to be made about what purposes. Further, the person or agency releasing data may be unable to control the use of data once released, so a certain amount of trust is necessary.

On the other hand, secondary analysts also need some consideration. A proposal to conduct specific analyses should not be rejected because it is philosophically objectionable to the primary researcher, or because the primary researcher does not believe that the line of investigation will be fruitful. In addition, in reporting results of secondary analyses, the analyst needs to provide some description of the data set to enable readers to evaluate the results of the secondary analyses. It is <u>not</u> an acceptable solution to report that an unspecified data set was used. Such a procedure requires implicit trust that the data set was faultless, or that the researchers were aware of and took into proper account all of the problems of the data set. Further, the person(s) conducting secondary analyses of data sets should not have to wait for years for access to a specific data set. Research questions as well as data collection procedures

and research methodologies become obsolete and, for some purposes, data collected five years ago are inappropriate.

A broader problem in the release of data sets involves letting people know that data sets are available. In some instances widespread announcements have been made, such as the recent announcement that data are available from the National Assessment of Educational Progress. However, in many other instances it is difficult to determine whether a particular data set is available. A researcher may have little or no knowledge of whether data sets exist which are appropriate for the consideration of certain types of questions. Informal networks of information about data sets aide those who are part of the networks but exclude others. While we have not resolved all of the issues about release of BTES data, we have had a few requests for data, all from persons with direct or indirect involvement in the research effort.

Persons struggling with these issues at the present time are typically struggling in the dark. It is apparent that a number of data sets have been subjected to analysis by persons other than the data collectors, yet there is very little discussion in the professional literature about the arrangements for such transfer of data. I have checked with a few friends who have been involved in such exchanges, but short of personal communication, there is little discussion of the issues involved and appropriate methods to protect everyone involved in such a release of data. We at the Commission have considered a number of options, each of which provides protection for one or more of the groups

of persons involved in the BTES. However, each of these options has one or more major drawbacks which makes it unattractive. We seem to fluctuate between trying to provide failsafe protection to the persons involved in the collection of data and trusting that anyone who asks for data will protect the interests of everyone involved.

Documentation of Data

When data sets are not adequately documented, they are of limited use to secondary analysts. The secondary researcher is compelled to seek answers to numerous questions from the primary researcher, and often to rely on the memory of the person(s) who collected the data.

Most proposals which involve the collection of data do not include a plan or budget for documentation of the data set. One of the problems in release of BTES data is the lack of adequate documentation of the data set and the procedures followed to obtain the data set, including collection, "cleaning", and any combining of data. In fact, at the moment, there is not agreement about what constitutes adequate documentation of a data set.

Next Steps

A number of next steps are necessary if we are to move to a more coherent use of data sets for additional analyses. Given the present cost of, as well as the number of present and predictable restrictions on, the collection data, the state of educational research would be well served by a move to more extensive use of data sets for further analyses.

Preparation for Release of Data

Some steps can and need to be taken in the early stages of any research effort involving the collection of data to insure that data will be available to other researchers for further analyses.

At the start of a research effort, agreement should be reached concerning the persons who will make decisions about what data will be released and for what purposes, as well as the procedures and criteria to be used. If not clearly stated at the start of a project, these questions may lead to a number of conflicts and misunderstandings.

Procedures for obtaining the informed consent of participants must allow for the release of data for further analysis to insure that data can be released later to other researchers. Thought must be given to the phrasing of descriptive materials about the research effort and to the types of commitments to be made to participants at the time of initial contact. The process of contacting participants at a later time to obtain such consent is difficult, costly, and not always possible.

The budget and work plan for any data collection effort should provide for documentation of the data set. If a copy of the data set is to be made available to the funding agency, a copy of the documentation should also be required. The work of documenting the data set should occur as the data are being collected and prepared for analysis, to the extent possible. Any after-the-fact documentation increases the possibility that the documentation will be inaccurate or incomplete. While it

may be easy to describe the instruments used to collect data it may be less easy to describe any events during the data collection which may have affected the accuracy of the collected data.

Discussion and development of a general consensus about documentation would provide guidance, or at least clarification, of various documentation issues. It is particularly difficult for persons immersed in the process of data collection to determine that their documentation is adequate, since they may read more information into the written documentation than is actually there. Further, persons collecting data for one purpose may have difficulty envisioning other uses for the data and the resultant documentation needs of other users. General agreement about the types of documentation needed might serve to limit some of these potential problems. For instance, everyone might agree that documentation should include descriptions of data collection instruments, data collection schedules, and procedures for training the persons collecting the data. Any unusual events which occurred during the collection of data should also be recorded. However, the definition of "unusual events" may not be clear. If the data set includes observation data collected in public school classrooms, than any number of events might affect the data, from a fire in the building to a threatened teacher strike to a shift in the instructional organization of classrooms. The persons collecting the data may or may not have information about any of these events, and the events may have different impacts on the observational data. A thorough discussion of the extent and types of documentation would potentially throw some light on these problems.

Release of Data

In light of the varied criteria used to decide what types of data should be released for further analysis, a general discussion is again appropriate. Ideally, a set of general guidelines or criteria, arising from an extensive discussion, might serve as a starting point for decisions about any particular data set. The guidelines would not, of course, be binding upon any researcher or research agency, since it is probably not possible to develop guidelines which are appropriate in all situations. Even without the development of guidelines, such a discussion would result in consideration and clearer statement of the problems associated with release of data for further analysis.

Another group of problems which would be illuminated and possibly minimized by extended discussion is the cluster of problems related to provision of data sets to secondary analysts while protecting the rights and commitments of participants and original researchers.

The discussions might address potential criteria for determining what data to release. Procedures might be discussed for determining when to release data and under what circumstances. Should the primary researcher(s) have some period of time in which to conduct analyses and report the results of the analyses? Should the primary researcher and/or the funding agency have the right to stop or delay release of reports of any analyses? Should the secondary analyst be free to conduct any analyses, provided that reports of results do not identify the source of the data?

Researchers and persons in funding agencies should give some thought to procedures for release of data sets. Among the questions to be considered are: ways to provide knowledge about data sets; ways to determine that a particular data set is appropriate for a specific purpose; procedures for review of requests for data; procedures for providing data sets and documtation and for payment of costs involved.

A number of problems, not adequately considered here, relate to the process of determining that a particular data set is appropriate to answer specific questions of interest. When a secondary analyst has a specific area of interest, or research question, in mind, there may be a number of data sets which potentially would be appropriate. The process of reviewing the data sets and determining which, if any, of them are appropriate may require the secondary analyst to review in detail the descriptions of a number of data sets to locate the one or more appropriate data sets. It may be that the absence of data about one variable will make the data set unusable for some purposes.

Given the present lack of adequate procedures for obtaining data for further analysis, I hesitate to even suggest that funding agencies consider whether data collection efforts are necessary for a particular research effort. However, at some time in the future it may be appropriate to require that any proposal for funding of data collection efforts provide an assurance that no existing, available data set contains the appropriate or required data and that a data collection effort is therefore necessary.

Data Bank

One approach to the problem of publicizing sources of data for further analysis is to establish a center responsible for collecting and cataloguing information about data sets. Such a center would provide 1) a source for persons seeking data sets for analysis and 2) a means whereby persons or agencies with data sets can make them available to other researchers.

There are a number of roles that such a center might play. The center might obtain and catalogue brief descriptions of data sets, directing interested researchers to the agency which collected or is storing the data set. At the other end of the continuum, the center might serve as a repository for data sets; a task that might involve development of detailed descriptions of data sets, review of proposals to determine whether any data sets in the repository are adequate to meet the proposed data needs, review of proposals to determine whether data should be provided, and duplication of data with appropriate documentation.

One role that a center could serve is that of catalyst in discussions of questions concerning release of data for further analysis. There are several points of view which need to be considered in any discussion of these issues, and there may be several steps necessary to adequately consider all of the viewpoints. However, it does seem essential that we begin to address issues of what data should be made available, what criteria should be used to determine the data that should be released, what con-

stitutes adequate documentation of data, and what constitutes adequate protection for the various involved persons and agencies.

Whoever initiates the discussion, it is time to begin.